ABSTRACT OF THE DISCLOSURE

An improved load balancing method, system, and computer program product handles popular object requests using a front-end cache, and hashing is applied only to the requests in the stream that were not handled by the front-end cache. A cache (e.g., a web proxy cache) is placed in front of a Level 7 switch, such that the cache services the popular requests from the cache based on the content of the request (e.g., based on the portion of an HTTP request following the domain name). The remaining requests are hashed and then routed to the back-end server. This allows the requests that make it past the cache to still be routed to the back-end server and take advantage of the efficiencies provided therefrom.

Preferably, a Level 4 switch is placed in front of a plurality of web proxy caches, each of which are in turn placed in front of a respective Level 7 switch, each of which are connected to a respective server farm, so that incoming web requests are handled on a round robin basis (or other SLB technique) before being sent to the cache, thus improving the throughput from the server farms to the requesting clients.

M:\MSimpson\Clients\IBM\24840 USA\Patent Office\RSW920010045US1-Final-WPerfect.wpd